Experimental UD Treebank for Mansi

Csilla Horváth
University of Helsinki, Finland / University of Tromsø, Norway

Key words: indigenous language, endangered language, low-resource language, Siberian language, UD database

1. Aim of the project

The aim of present project is to create an experimental Universal Dependency Treebank for the Mansi language.

2. The Mansi language

Mansi is a severely endangered Uralic language spoken in Western-Siberia. According to the 2020 census data of the Russian Federation, there are 12,228 Mansi in Russia (Census RF 2020 5/1). A total of 1,346 people stated that they spoke Mansi (Census RF 2020 5/4), and 1,236 of them were of Mansi ethnicity (Census RF 2020 5/19). Mansi is used in both spoken and written form. It is spoken most often in private life with relatives and childhood friends. The written language is used primarily in the monthly newspaper Luima Seripos.

Mansi is an agglutinating, transitive-accusative language low in fusion and flexion. The basic constituent order in a neutral clause is S(O)V. The finite verb is in clause-final position in both active and passive sentences. Nouns in Mansi are inflected for number, possession, and case. There are three numbers: singular, dual, and plural. There are six cases: nominative, lative, locative, ablative, instrumental, and translative. The translative case is generally not used in the dual and plural. Verbs have the inflectional categories of subject person, subject and object number, tense, mood, and voice (c.f. Virtanen & Horváth 2023).

3. Attempts to create language technology tools for Mansi

Only a couple of attempts are known that aimed to create language technology tools for Mansi. In 2009, morphological analysers were created on the materials of Mansi scratch grammars and text collections (c.f. Fejes and Novák 2010), written with Finno-Ugric Transcription. In 2013, the creation Mansi-Russian and Mansi-Hungarian dictionaries, a Mansi corpus of 500k tokens, Mansi morphological analyser and Mansi Word-Net was started (c.f. Horváth et al 2015). In 2020, the beta version of an Android-based Mansi language learning tool was created with Hungarian as its language pair (c.f. Bobály et al 2020). In 2023, the creation of a Mansi spellchecker started[1]. Although a small dependency treebank of Mansi was also planned to be created, it never actually took place.

4. Resources related to the project

An independent project aiming to archive all the available Mansi materials published with Cyrillic transcription between 1937 and 2020 has been started in 2021 (c.f. Horváth 2021). The corpus is going to contain the folkloric texts (volumes of recently collected legends, folktales, stories), the literary texts (children's literature, contemporary prose), the academic texts

[1] https://divvun.org/proofing/online-speller.html

(papers, non-fiction publications for children), the translation literature (political pamphlets, socialist agitating publications, socialist literature), the Bible translations (Gospels, other biblical fragments) published in Mansi, as well as all the available materials from the Mansi press. Achieving the aims of the project, the largest corpus of Mansi texts is going to be created. The expected size of the corpus is 2-2.5 million tokens.

5. Project plan

The realisation of the project plan is in initial phase. The necessary Mansi materials have been collected and have been digitalised. The OCR adaptation of the texts, and the manual correction of the texts after processing are still going on. For the summer school, I am choosing 150 Mansi sentences for morphological annotation: 100 sentences from the Mansi press and 50 sentences from Mansi literature. The sentences are going to be annotated manually.

During the training school, I would like to get familiar with the principles of UD Treebank building, in order to create the first, experimental dependency corpus for Mansi in harmony with the Universal Dependencies project. This knowledge would prove useful when the written Mansi corpus and the Mansi morphological analyser are released, as it would enable the creation of a larger Mansi UD Treebank database as well.

Bibliography

Bobály, Gábor, Horváth, Csilla and Vincze, Veronika. 2020. apPILcation: an Android-based tool for learning Mansi. In: A Pirinen, Tommi; M. Tyers, Francis; Rießler, Michael [eds.] Proceedings of the 6th International Workshop on Computational Linguistics of Uralic Languages. Stroudsburg: Association for Computational Linguistics, 48-55.

Census RF. 2020. 5/1, Национальный состав населения https://rosstat.gov.ru/storage/mediabank/Tom5_tab1_VPN-2020.xlsx Accessed 1 April 2023.

Census RF. 2020. 5/4, Владение языками и использование языков населением. https://rosstat.gov.ru/storage/mediabank/Tom5_tab4_VPN-2020.xlsx. Accessed 1 April 2023.

Census RF. 2020. 5/19, Владение языками коренных малочисленных народов Российской Федерации. https://rosstat.gov.ru/storage/mediabank/Tom5_tab19_VPN-2020.xlsx. Accessed 1 April 2023.

Fejes, László and Novák, Attila. 2010. Obi-ugor morfológiai elemzők és korpuszok. In: Tanács, Attila and Vincze, Veronika [eds.]: VII. Magyar Számítógépes Nyelvészeti Konferencia. Szeged: Szegedi Tudományegyetem, 284-291.

Horváth, Csilla. 2021. “Mansi corPŌS”: Archiving Mansi Written Materials Published between 1937 and 2020. In: Jared, Desjardins [ed.]: Proceedings of the 4th Workshop on Computational Methods for Endangered Languages Colorado: University of Colorado, 16-19.

Horváth, Csilla, Szilágyi, Norbert, Veronika, Vincze and Nagy, Ágoston. 2017. Language technology resources and tools for Mansi: An overview. In: Proceedings of the 3rd International Workshop on Computational Linguistics for Uralic Languages. 46-55.

Virtanen, Susanna and Horváth, Csilla. 2023. Mansi. In Abondolo, Daniel [ed.]: The Uralic languages, 2nd edition, London: Routledge, 665-702.